TP 2: Análisis de Redes Sociales

October 12, 2022 Métodos Numéricos

Grupo 18

Integrante	LU	Correo electrónico
Vekselman, Natán	338/21	`natanvek11@gmail.com`
Arienti, Federico	316/21	`fa.arienti@gmail.com`

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2610 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

Tel/Fax: (++54 +11) 4576-3300

http://www.exactas.uba.ar

RESUMEN

La descomposición de matrices en autovectores y autovalores aparece en una variedad de aplicaciones donde importa caracterizar el comportamiento de un sistema: en el reconocimiento de imágenes, en el análisis de estabilidad de cuerpos rotantes, en el análisis de riesgo de mercado y en el análisis de redes —por nombrar algunas—. Desde un punto de vista geométrico, se puede considerar a los autovectores como los 'ejes' de una transformación lineal, en tanto representan una dirección invariante a la transformación, y a los autovalores como los factores por los que esas direcciones se comprimen, estiran o invierten.

En este trabajo implentaremos un método para el cálculo de autovectores y autovalores en matrices con autovalores en módulo diferentes —descontando multiplicidad—, como pueden ser las matrices simétricas definidas positivas, las matrices de conectividad o las matrices diagonalizables. El mismo se conoce como *el método de la potencia con deflación*.

A su vez, presentaremos dos aplicaciones concretas de los autovalores y autovectores en el análisis de redes: la medición de centralidad de autovector y corte mínimo en la red del *'Club de Karate'* [citar] , y la estimación de una *ego-network* [citar] en Facebook, por medio de la construcción de una matriz de similaridad.

Palabras clave: *método de la potencia, deflación de Hotelling, centralidad de autovector, conectividad algebráica, análisis de componentes principales.*

CONTENIDOS

1. Método de la potencia con deflación

1.1. Introducción teórica. El método de la potencia con deflación permite aproximar un subconjunto de los autovalores y autovectores asociados a una matriz. Si la misma satisface que todos sus autovalores son diferentes en módulo, entonces permite aproximar el conjunto entero. El método se compone en dos partes:

Método de la potencia: el método de la potencia, *Power method* o *Power iteration*, es una técnica iterativa para aproximar el autovector asociado al autovalor de módulo máximo de una matriz cuadrada que satisfaga esta característica —es decir, tenga un autovalor dominante—, a partir de la aplicación de sucesivos productos matriciales, descriptos por la siguiente relación de recurrencia:

$$b_0 \text{ es un vector aleatorio} : ||b_0|| = 1$$

$$b_{k+1} = \frac{\mathbf{A} b_k}{||\mathbf{A} b_k||} \tag{1}$$

donde $||\cdot||$ es una norma vectorial.

Se puede demostrar [citar] que, bajo las condiciones descriptas, si b_0 no es ortogonal al autovector asociado al autovalor dominante en módulo de $\mathbf{A}$, b_k convergerá a éste. Lo que es más, se podrá aproximar el autovalor dominante por medio del coeficiente de Rayleigh [citar] :

$$\lambda_{max} = \frac{b_k^t \ \mathbf{A} \ b_k}{b_k^t \ b_k} \tag{2}$$

Método de la deflación: el método de la deflación, por su parte, corresponde a la transformación de la matriz inicial $\mathbf{A}$ por una matriz $\mathbf{B}$ que comparta sus autovalores salvo por el autovalor dominante que será anulado. Existen numerosos métodos de deflación, entre ellos la deflación de Hotelling [citar] y la deflación de Wielandt [citar] .

[completar con una explicación del método que usemos.]

1.2. Implementación. Procedemos a detallar una posible implementación para ambos métodos. Para no complicarnos con la representación de números complejos, nos restringiremos a matrices cuyos autovalores asociados sean reales. Definimos:

$$\textit{deflacion}: \ \textit{matriz}_{n\times n}\ \mathbf{A}\ \times\ \textit{nat}\ \mathrm{q}\ \times\ \textit{nat}\ \mathrm{k}\ \times\ \textit{real}\ \mathrm{t}\ \longrightarrow\ \textit{vector}_q\ \times\ \textit{matriz}_{n\times q}$$

$$\textit{potencia}: \ \textit{matriz}_{n\times n}\ \mathbf{A}\ \times\ \textit{nat}\ \mathrm{k}\ \times\ \textit{real}\ \mathrm{t}\ \longrightarrow\ \textit{real}\ \times\ \textit{vector}_n$$

donde n es un natural, $\mathbf{A}$ tiene q autovalores reales dominantes en módulo, $0 < q \leq n$, k representa el número máximo de iteraciones a realizar y $0 \leq t$ representa la tolerancia aceptable de error.

Proponemos el siguiente algoritmo para *deflacion*:

```
proc deflacion(in A: matriz<n, n>, in q: Nat, in k: Nat, in t: Real) {
    // TODO
}
```

Algoritmo 1. Pseudocódigo para el método de la deflación.

El mismo retornará un vector con los q autovalores de $\mathbf{A}$, ordenados descendientemente en módulo, y una matriz cuyas columnas corresponden, respectivamente, a los autovectores asociados a los q autovalores.

[explicar aspectos relevantes a la implementacion.]

potencia por su parte, retornará el autovalor de $\mathbf{A}$ máximo en módulo y su autovector asociado. Proponemos el siguiente algoritmo:

```
proc potencia(in A: matriz<n, n>, in k: Nat, in t: Real) {
    // TODO
}
```

Algoritmo 2. Pseudocódigo para el método de la potencia.

[explicar aspectos relevantes.]

1.3. **Evaluación cuantitativa.** Procederemos a evaluar nuestra implementación del método de la potencia con deflación en C++ acorde a los algoritmos propuestos.

Error relativo Medimos el error $|\mathbf{AV} - \mathbf{V\Lambda}|_1$ en función de la cantidad de iteraciones k para 300 instancias de matrices $\in \mathbb{R}^{25\times 25}$ generadas aleatoriamente, donde $\mathbf{V}$ y $\mathbf{\Lambda}$ representan —respectivamente— las matrices aproximadas de autovectores y autovalores de $\mathbf{A}$, tal que $\mathbf{AV}_i \approx \mathbf{\Lambda}_{ii}\mathbf{V}_i\ \forall i:\ 1 \ldots n$. En total, obtuvimos [XXX] mediciones[1].

[1] El script asociado se puede encontrar en *./experimentos/error_potencia.py*

Metodología. Se calculó $\mathbf{\Lambda}, \mathbf{V} = Deflacion(\mathbf{A},\ 25,\ k,\ 0)$ y se midió el error relativo para cada una de las matrices sobre cada valor de k en en el intervalo $(0, 1e6)$, de a saltos de $1e3$.

Cada caso se generó a través de uno de los siguientes tres procedimientos[2]:

1) *Matrices Diagonales*: Se generaron cien matrices diagonales $\mathbf{D}$ con veinticinco autovalores en el rango $[-1e6,\ 1e6]$ tal que ningun autovalor compartiera valor absoluto en módulo con ningún otro. Los mismos se generaron con el rng *PCG64* de numpy para evitar distribuciones particulares que pudieran influir en la variabilidad de los autovalores.
2) *Matrices Diagonalizables*: Se generaron cien matrices diagonalizables $\mathbf{A} = \mathbf{QDQ}^t$ donde cada matriz $\mathbf{D}$ se generó a partir de la metodología (1.) y $\mathbf{Q} = \mathbf{I} - 2uu^t$ se generó a partir de un vector aleatorio u —con el algoritmo random.rand() de numpy— tal que $||u||_2 = 1$.
3) *Matrices Simétricas Definidas Positivas*: Se generaron cien matrices simétricas definidas positivas de enteros. Para cada caso se generó una matriz aleatoria $\mathbf{B}$ con el algoritmo *random.randint()* de numpy y se procedió a definir la matriz $\mathbf{A} = \mathbf{BB}^t$.

Resultados. [experimento]

[2] Se utilizó un valor semilla para facilitar la reproductibilidad.

2. Análisis: Club de Karate

2.1. **Contexto.** La red del *Club de Karate* es parte de una investigación antropológica [citar] que estudió las relaciones 'políticas' entre los miembros de un club universitario. La misma se realizó durante el desarrollo de un conflicto que terminó por dividir al grupo. La red busca modelar el flujo de información entre sus integrantes por medio de la tripla ($\mathbf{V}$, $\mathbf{E}$, $\mathbf{C}$), donde $\mathbf{V}$ denota el conjunto de individuos, $\mathbf{E}$ refiere a un grafo no direccionado cuyos ejes representan los vínculos entre los miembros, y $\mathbf{C}$ define la fuerza de estas relaciones —lo que se podría pensar como ponderaciones sobre los ejes de $\mathbf{E}$—.

La investigación tuvo como eje central demostrar la capacidad del modelo para predecir la división del grupo por medio del algoritmo de *labeling* de 'flujo máximo - corte mínimo' de Ford y Fulkelson [citar] .

En este análisis utilizaremos sólo la representación matricial de $\mathbf{E}$ para evaluar la importancia de los distintos miembros en la red, y la matriz laplaciana asociada para evaluar el uso de autovectores como predictores de la división del grupo.

2.2. **Centralidad de Autovector.** La centralidad de autovector es una medida que se utiliza en el análisis de redes para medir la 'importancia' de los nodos que componen una red, relativa a la importancia de sus conexiones. Dada una matriz de conectividad $\mathbf{W}$, se define:

$$\lambda x = \mathbf{W}x \tag{3}$$

donde λ es el autovalor en módulo máximo de $\mathbf{W}$.

Intuitivamente, se puede pensar que la importancia de cada nodo es proporcional a la suma de las importancias de sus vecinos [citar] . Se puede demostrar [citar] que, dado las características de esta matriz, el autovector asociado siempre será positivo en coordenadas.

En tanto la red de *Club de Karate*, podemos ver que la aplicación del metodo de la potencia con deflación sobre la matriz de conectividad asociada al grafo $\mathbf{E}$ resulta en el siguiente autovector asociado al autovalor en módulo máximo:

[presentar vector.]

Vemos que el nodo '1' y el nodo '34' son los más 'centrales'. Esto no es casualidad, la red del *Club de Karate* está armada para tener a las dos figuras centrales del conflicto en cada extremo —el instructor de karate y el presidente del club—, para satisfacer la especificación del algoritmo de 'labeling' que utiliza.

2.3. **Autovectores de la matriz laplaciana.** La matriz laplaciana $\mathbf{L} = \mathbf{D} - \mathbf{W}$ —donde $\mathbf{D}$ es una matriz diagonal con elementos $d_{ii} = \sum_j w_{ij}$ y $\mathbf{W}$ es una matriz de conectividad—

sirve para medir distintas propiedades en una red. En particular, el mínimo autovalor en módulo no nulo —llamado de conectividad algebráica, o valor de Fiedler— permite establecer un criterio sobre el que particionar la red en dos. El autovector asociado a este autovalor designará la pertenencia de un nodo a una u otra partición acorde a su signo.

Procedemos a analizar qué autovector de la matriz laplaciana asociada a **E** permite predecir mejor la división que ocurrió en el *Club de Karate*. Para ello, medimos el valor absoluto de la correlación entre cada autovector y un vector que indica la división que ocurrió en el grupo.

[presentar solución.]

3. Análisis: Red ‘Ego’

3.1. **Contexto.** Una *ego-network* [citar] es una red compuesta por las amistades que existen entre los amigos de un individuo, el ‘ego’. Estas redes son centrales para aplicaciones como Facebook, Google+ o Twitter. En particular, dada una red ‘ego’, resulta de interés poder identificar los circulos sociales —conjuntos, disjuntos y anidados— a los que pertenece un usuario. En [citar] se propone un método de aprendizaje no supervisado para lograr inferirlos, que se nutre de la siguiente información: un grafo **E** —la red—, donde se espera que exista una correlación fuerte entre un círculo y la densidad de conexiones entre los nodos que lo componen; y un conjunto de atributos **C**, para cada nodo, donde se espera una correlación entre un círculo y la similaridad de los atributos de los nodos que lo componen—.

En este análisis estimaremos **E**, la red ‘ego’ original, por medio de la construcción de una matriz de similaridad que utilice los atributos definidos en **C**. También, buscaremos reducir su dimensionalidad por medio del análisis de componentes principales.

3.2. **Matriz de similaridad.**

3.3. **Comparación con la red original.**

3.4. **Optimización.**

3.5. **PCA.**

4. Conclusiones

5. Apéndice